Introduction to Digital Communication: Modulation

An Online Continuing Education Course for Engineers

Course Number: E-1014

Credit: 1 Hour / 1 PDH / 1 CPD

Introduction to Digital Communication: Modulation

Carl B. Dietrich, Jr., PhD, P.E.

Introduction

Digital communication has become ubiquitous in our society. Like much of the world's population, most engineers are regular users of digital services such as the Internet, cellular and land-line telephone networks, and digital television. All of these services depend on digital modulation techniques to enable efficient transmission and reception of signals. In addition, our work may involve digital communication in several ways. While some engineers design new digital communications systems, many others need to specify communication systems for use in factories, mines, "smart grid" approaches to power transmission and distribution, and other applications including telemetry and control. To do so, it is helpful to understand the underlying technologies of these systems.

This course provides an introduction to one of many technologies that make digital communication possible. Digital modulation and the inverse process of demodulation, transform data or digitized voice or video into a form that can be transmitted and received over a channel and then transform it back to its original form. The course emphasizes radio transmission, but modulation and demodulation are equally important in systems that use other transmission media, such as radio frequency or fiber optic cable links.

The course begins with a brief overview of digital communications systems. This is followed by a discussion of some of the challenges that must be overcome for effective digital communication, such as noise and interference. Next the concept of channel capacity is introduced, including Shannon's expression for capacity of a noisy communication channel and a straightforward example of how to calculate this upper limit on error-free data transmission rate. Digital modulation and demodulation are discussed, starting with their goals and concepts necessary to understand the topic. Specific types of digital modulation are described as are important implementation issues, techniques used to achieve reliable communications, and visualization of digital modulation through constellation plots. Finally, a list of common digital modulation types and their advantages and disadvantages is given.

A brief course cannot be comprehensive. Many topics are omitted or only mentioned briefly, including mapping of bits to transmitted symbols and use of grey coding, pulse shaping, eye diagrams, and details of synchronization and demodulation including matched filtering, symbol decision statistics, symbol errors and bit errors. References [1] and [3] are recommended for readers seeking further detail.

Why Digital Communications?

Digital communication is necessary between computers, which use and exchange information in digital form. There are also compelling reasons for using digital modulation to transmit analog signals such as voice or music. These include:

1. Nearly error free transmission (high fidelity) provided that the signal-to-noise ratio is above a threshold level
2. Ability to compress data and trade off fidelity and spectrum efficiency
3. Compatibility with efficient and robust multiplexing and multiple access schemes that are not possible with analog modulation
4. Ability to transmit analog and digital information over the same network using identical or compatible communication standard

Functions, units and terms used in this course

Logarithms: The logarithm with base a of x or $\log_a x$ is defined as follows

if $y=\log_a x$ then $x=a^y$.

In converting ratios to decibels (see below) we use log base10. In that case, $y=\log_{10} x$ indicates that $x=10^y$.

Capacity of a communications channel for binary data (in which each digit has two possible values) is calculated using log base 2 as described later in this course.

Decibels (dB): a logarithmic unit used to express the ratio of two power or voltage levels or to express power or voltage relative to a reference level. Decibels are used in communications engineering because they allow very large or small ratios to be expressed in a compact form. Use of decibels converts very small or large numbers to more manageable negative and positive numbers. This combined with properties of the logarithm[1] result in simplified computation, allowing quick mental or "back of the envelope" calculations.

The ratio of two power levels can be expressed in dB as $(P_1/P_2)[dB]=10\log_{10}(P_1/P_2)$

Because power is proportional to the square of voltage, the ratio of two voltages can be expressed in dB as $(V_1/V_2)[dB]=10\log_{10}(V_1^2/V_2^2)= 10\log_{10}(V_1/V_2)^2=20\log_{10}(V_1/V_2)$

Power can also be expressed in dBm (decibels relative to 1mW): P[dBm] = 10log10(P/1mW)

Bandwidth (typically represented by W or B): the range of frequencies occupied by a communications signal. For example, a signal that occupies the frequencies between 101.2 MHz and 101.4 MHz is said to have a bandwidth of 0.2 MHz or 200kHz.

[1] Use of logarithmic units such as dB allows replacement of multiplication with addition, division with subtraction, and exponentiation with multiplication, due to the following properties of the logarithm: $\log(xy)=\log x+\log y$, $\log(x/y)=\log x-\log y$, $\log(x^y)=y\log x$

Digital Communication Systems and Networks

Digital communication and digital modulation predate the use of electricity by humans: early examples include smoke signals. In the simplest form, presence or absence of smoke could have two distinct meanings. U.S. readers are likely to have heard the story of Paul Revere and the use of lanterns to signal the route taken by British forces ("one if by land, two if by sea,") during the American Revolutionary War. Early electrical communication was digital; Morse code uses combinations of four basic symbols, a short pulse represented by a dot (**.**), a longer pulse represented by a dash (**–**), a short space that separates letters, and a longer space that separates words, to allow transmission of more complex information. Telegraphy was the earliest practical means of communication by radio and is still used. It is very reliable but not extremely fast, and is effective under channel conditions in which other, potentially faster, types of communication are not. This tradeoff between speed and reliability is fundamental, and it can be addressed by choosing from among the many available types of digital modulation.

<u>Modulation and Demodulation in the context of communication systems and networks</u>

Today, many of the digital communication systems we use are part of networks of computers and other devices. The functionality of a communication network can be represented by and organized into layers as in the open systems interconnect (OSI) stack shown in Fig. 1 [1]. In this context, modulation and demodulation are part of the Physical layer, the lowest layer in the OSI model that enables communication of data. Higher layers enable devices to share a medium or channel, provide routing of data to specific devices, and ultimately allow users to interact with the network through applications such as a web browser or voice over Internet protocol (VoIP) client, but all these layers depend on the performance of the physical layer, including modulation and demodulation.

OSI Layer
7. Application
6. Presentation
5. Session
4. Transport
3. Network
2. Data link
1. Physical

Figure 1: Open Systems Interconnection (OSI) reference model or stack representing different layers of functionality in digital communication networks (adapted from [1], Fig. 12.6-2, p. 538).

<u>Components of a communication system</u>

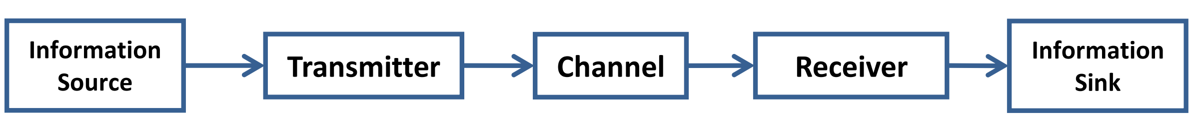


Figure 2: Block diagram of a basic communication system (adapted from [1], Fig. 1.1-1, p. 3)

A communication system in its simplest form (see Fig. 2) consists of a transmitter, which sends information from some source (e.g., a microphone, camera, or computer) over a channel, and a receiver that has the function of recovering the original signal or a close approximation in a form that is intelligible to a human or machine, and delivering it to a destination or information sink [1]. Simplex communication involves a single transmitter. For simplicity, only a simplex or unidirectional link is shown. Point-to-point communication with a single receiver and broadcasting to multiple receivers are both possible. Duplex communication involves a collocated transmitter and receiver at each end of the communication link. In half-duplex communications, transmission occurs in one direction at a time, and in full-duplex communications, simultaneous transmission in both directions is possible. The communications channel typically degrades the signal or transmitted information by introducing distortion and by reducing the amplitude of the signal. Degradation of the signal due to the presence of other interfering signals or noise can also be modeled as part of the channel. These effects may make it difficult or impossible for the receiver to recover an exact copy of the transmitted information. Also, every channel has a finite capacity, the maximum rate at which information could possibly be transmitted and received over the channel [2].

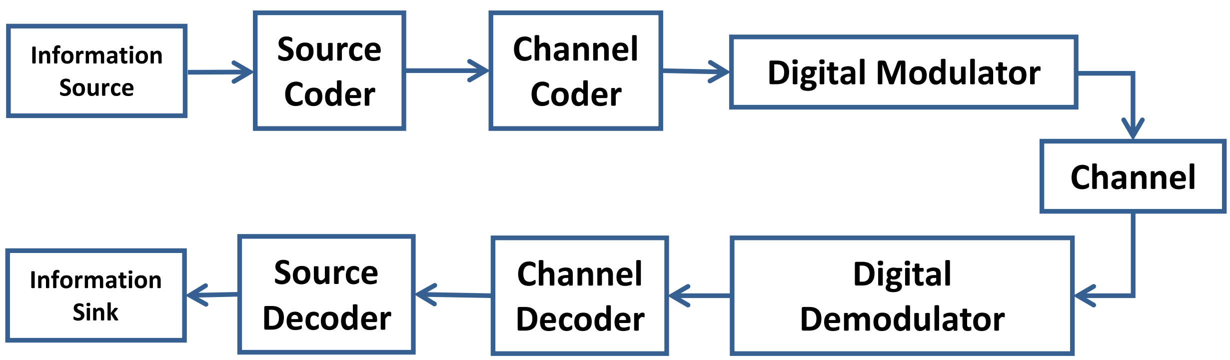
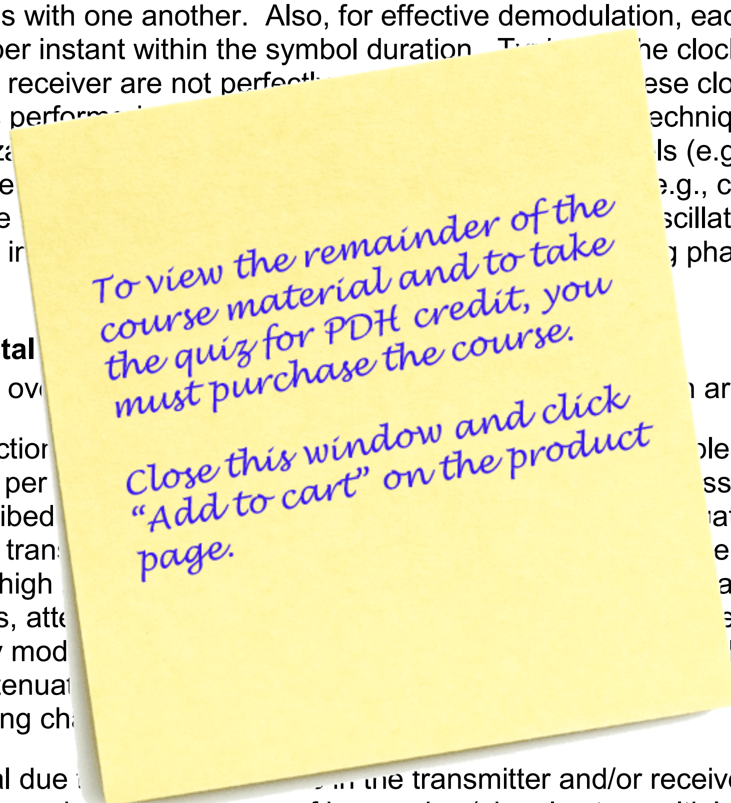


Figure 3: Digital Communications System (adapted from [3], Fig. 2.1.1, p. 65)

<u>Digital communication system</u>

Figure 3 shows a digital communication link, shown as simplex or unidirectional for simplicity. In a digital communication system, the source may be an analog source such as a microphone, camera, thermocouple, or other transducer, in which case the information from the source is converted to a digital form, but could also be an inherently digital source such as a computer transmitting data. A source coder compresses the digital information from the source to reduce the amount of information that must be transmitted over the channel. An example of this is the MP3 compression standard commonly used to compress digital audio files. Some types of compression degrade the signal, but in a way that is controlled by the system designer. The output of the source coder is passed through a channel coder, which introduces redundancy to the information in order to allow the receiver to detect and possibly correct errors in the received data. A modulator converts the data from the output of the channel coder into a signal that can be transmitted over the communications channel, and the channel introduces some degradation to the transmitted signal, which may result in errors in the received data. In the receiver, a demodulator converts the transmitted signal back into a digital form, and the channel decoder and source decoder perform the inverse operations of the channel coder and source encoder, respectively. Finally, the recovered signal is converted back to analog form if needed and then passed to the final destination of the signal (called an information sink). If the source is digital and the transmission rate is within the channel capacity, it is possible that bit errors either will not occur or can be corrected by channel coding. In this case, the original signal may be recovered perfectly despite the effects of the channel.

Timing and synchronization are very important in digital communications. Because the transmitted signal represents a sequence of numbers as a sequence of transmitted symbols, it is necessary to know when each symbol begins and ends in order to avoid confusing consecutive symbols with one another. Also, for effective demodulation, each symbol must be sampled at the proper instant within the symbol duration. Ty... he clock at the transmitter and the clock at the receiver are not perfect... ese clocks determine the timing of operations perform... echniques are needed to maintain synchroniz... ls (e.g. determining when a frame or sequence ... e.g., compensating for frequency offset due ... scillator frequencies, or due to Doppler shift in ... phase lock).



Challenges of Digital ...

Digital transmission ov... are described here:

Attenuation or reductio... ples, attenuation may be expressed in dB per ... ss communication, attenuation is described ... uation is proportional to the square of the tran... e transmitted energy. This results in very high ... ation. In land-mobile and indoor channels, atte... e Friis transmission equation, and many mod... hese channels. Time variation in attenuat... /or receiver as well as due to time varying ch...

Distortion of a signal due ... in the transmitter and/or receiver is also a consideration. This can include presence of harmonics (signals at a multiple of the desired signal frequency) as well as intermodulation products that occur due to interaction of multiple signals.

Interference due to other signals transmitted on the same frequency or on frequencies close to that of the desired signal can be a major source of signal degradation in a shared medium such as a radio channel.

Multipath effects occur when a signal arrives at the receiver via multiple paths, and is a major problem in mobile and indoor radio channels. The multiple copies of the signal can either cancel or reinforce each other, depending on their phase relationships. The phase relationship between or among the multiple received signals in turn depends on location and/or time. Thus the received signal may fade in and out as a receiver or transmitter moves through an environment in which multiple signal paths are present. An additional multipath effect relevant to digitally modulated signals is *inter-symbol interference*. Inter-symbol interference occurs